High Throughput Distributed Computing - 3

Stephen Wolbers, Fermilab
Heidi Schellman, Northwestern U.

Outline – Lecture 3

- **Trends in Computing**
- **Future HEP experiments**
  - Tevatron experiments
  - LHC
  - Other
- **Technology**
  - Commodity computing/New Types of Farms
  - GRID
  - Disk Farms

In Silica Fertilization

All Science Is Computer Science

By GEORGE JOHNSON

EXCEPT for the fact that everything, including DNA and proteins, is made from quarks, particle physics and biology don't seem to have a lot in common. One science uses mammoth particle accelerators to explore the subatomic world; the other uses petri dishes, centrifuges and other laboratory paraphernalia to study the chemistry of life. But there is one tool both have come to find indispensable: supercomputers powerful enough to sift through piles of data that would crush the unaided mind.

Last month both physicists and biologists made announcements that challenged the tenets of their fields. Though different in every other way, both discoveries relied on the kind of intense computer power that would have been impossible to marshal just a few years ago. In fact, as research on so many fronts is becoming increasingly dependent on computation, all science, it seems, is becoming computer science.

"Physics is almost entirely computational now," said Thomas B. Kepler, vice president for academic affairs at the Santa Fe Institute, a multidisciplinary research center in New Mexico. "Nobody would dream of doing these big accelerator experiments without a tremendous amount of computer power to analyze the data."

New York Times,
Sunday, March 25, 2001

Trends in Computing

- It is expected that all computing resources will continue to become cheaper and faster, though not necessarily faster than the computing problems we are trying to solve.
- There are some worries about a mismatch of CPU speed and input/output performance. This can be caused by problems with:
  - Memory speed/bandwidth.
  - Disk I/O.
  - Bus speed.
  - LAN performance.
  - WAN performance.

Computing Trends

- **Nevertheless, it is fully expected that the substantial and exponential increases in performance will continue for the foreseeable future.**
  - CPU
  - Disk
  - Memory
  - LAN/WAN
  - Mass Storage

Moore's Law

http://sunsite.informatik.rwth-aachen.de/jargon300/Moore_sLaw.html

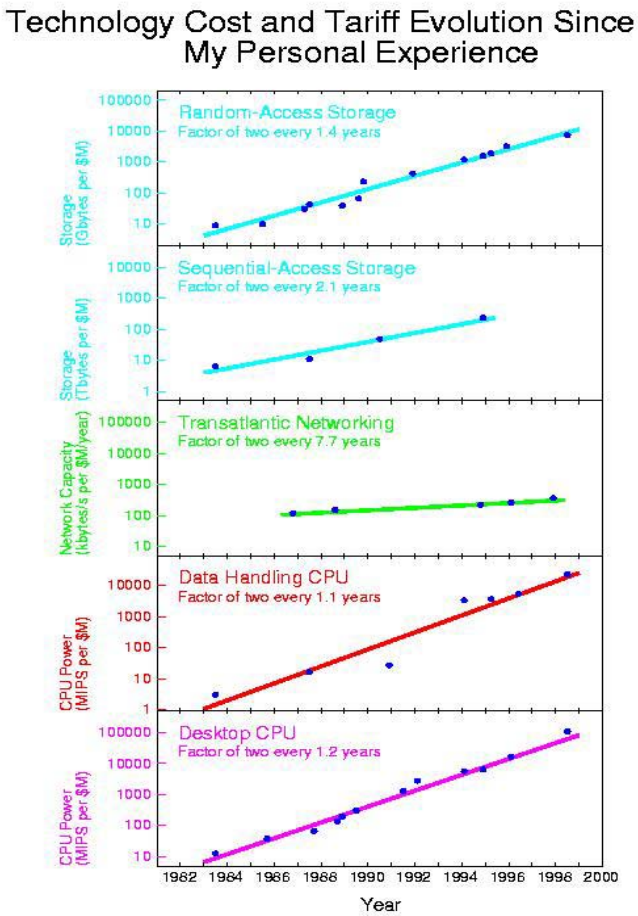
- density of silicon integrated circuits has closely followed the curve (bits per square inch) = 2^((t - 1962)) where t is time in years; that is, the amount of information storable on a given amount of silicon has roughly doubled every year since the technology was invented. See also Parkinson's Law of Data.

Parkinson's Law of Data

http://sunsite.informatik.rwth-aachen.de/jargon300/Parkinson_sLawofData.html

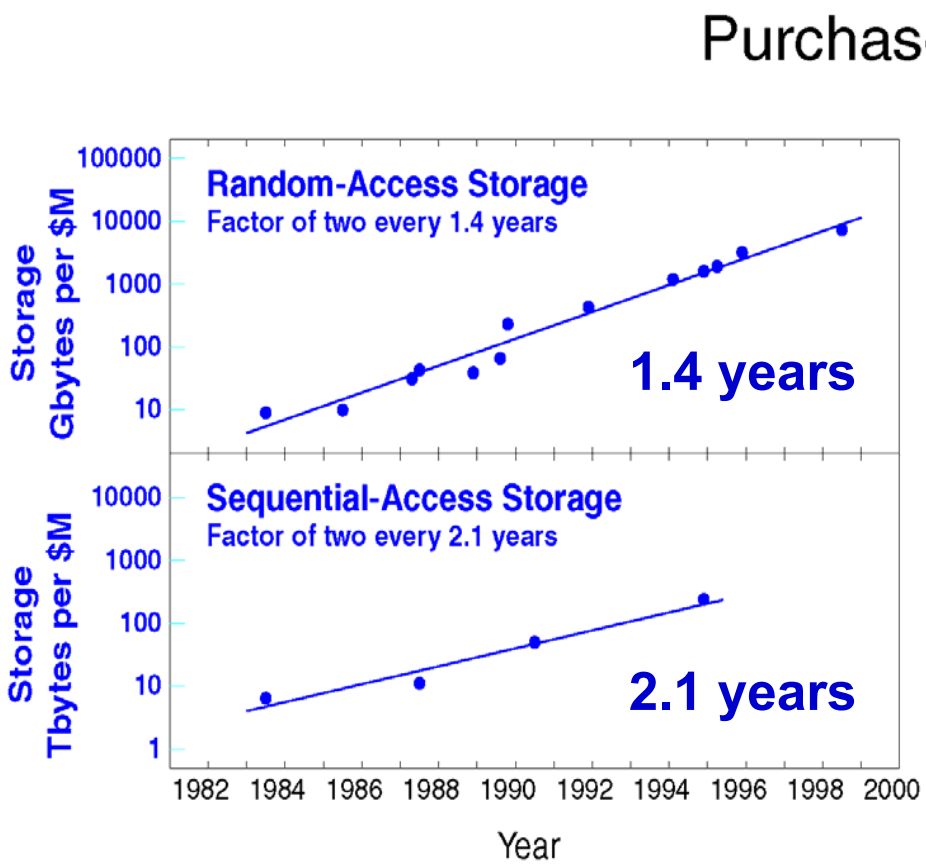
- "Data expands to fill the space available for storage"; buying more memory encourages the use of more memory-intensive techniques. It has been observed over the last 10 years that the memory usage of evolving systems tends to double roughly once every 18 months. Fortunately, memory density available for constant dollars also tends to double about once every 12 months (see Moore's Law); unfortunately, the laws of physics guarantee that the latter cannot continue indefinitely.

General Trends

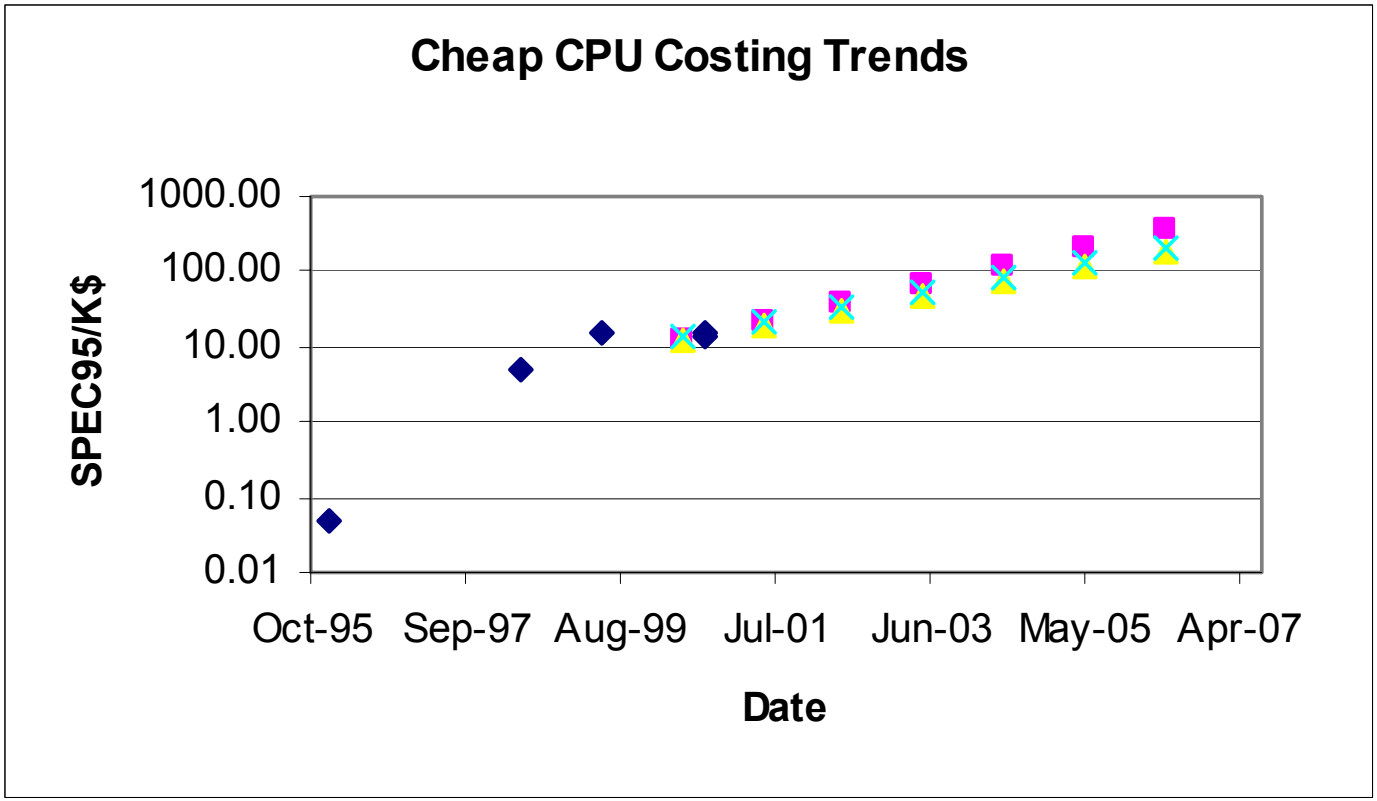


Hardware Cost Estimates

Paul Avery

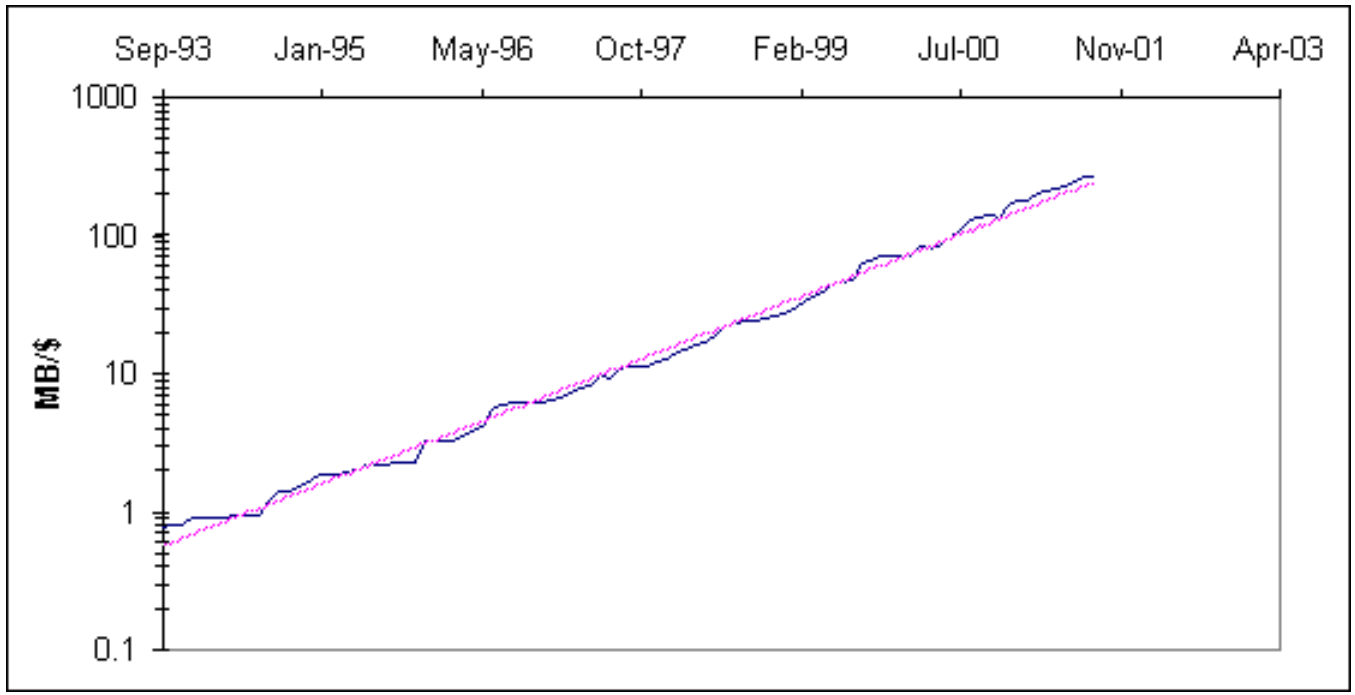


CPU Speed and price performance



Disk Size, Performance and Cost

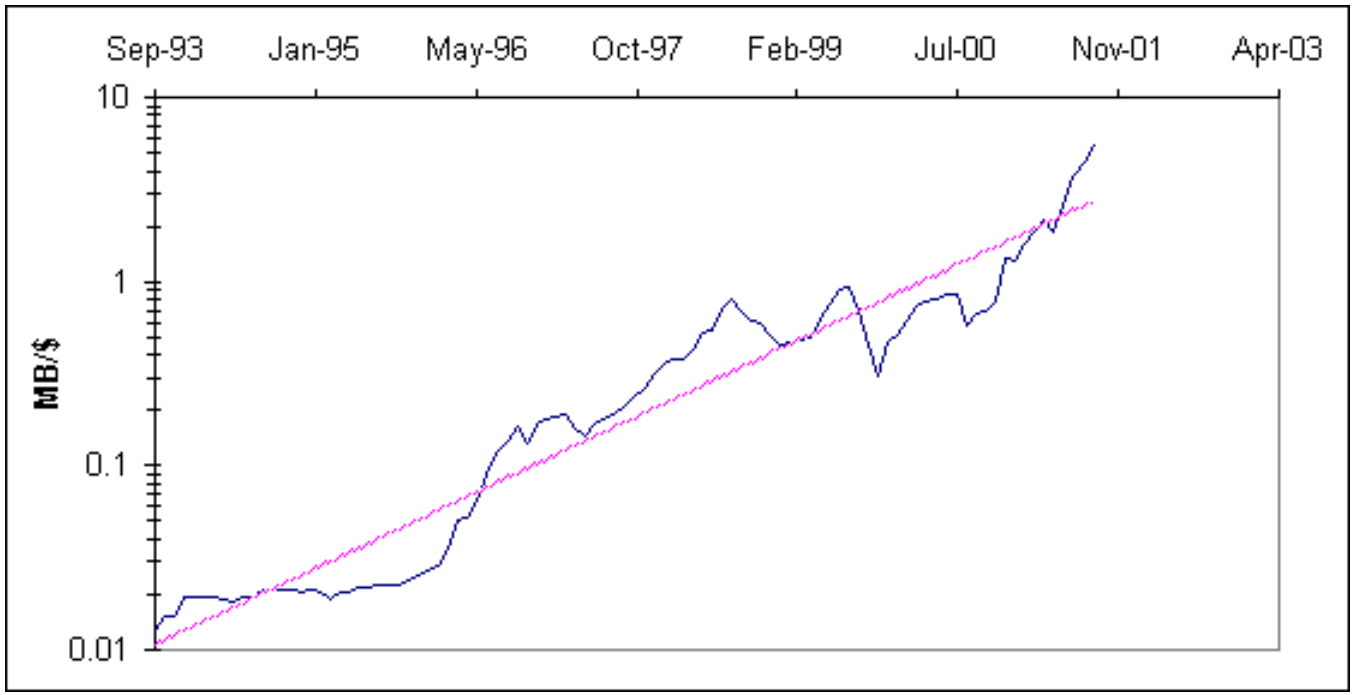
http://eame.ethics.ubc.ca/users/rikblok/ComputingTrends/



Doubling time = 11.0 +- 0.1 months

Memory size and cost

http://eame.ethics.ubc.ca/users/rikblok/ComputingTrends/

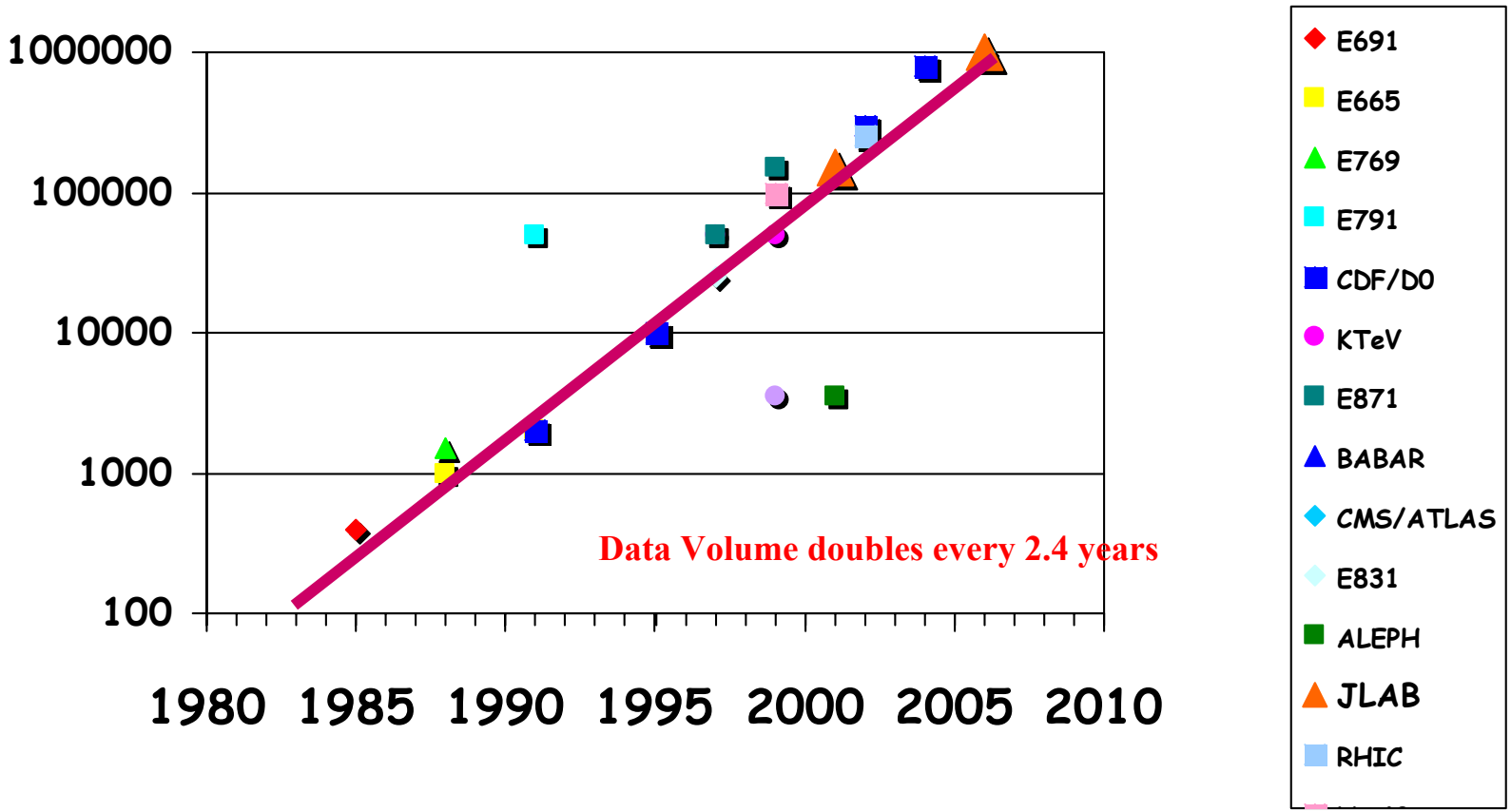


Doubling time = 12.0 +- 0.3 months

Worries/Warnings

- Matching of Processing speed, compiler performance, cache size and speed, memory size and speed, disk size and speed, and network size and speed is not guaranteed!
- BaBar luminosity is expected to grow at a rate which exceeds Moore's law (www.ihep.ac.cn/~chep01/presentation/4-021.pdf)
- This may be true of other experiments or in comparing future experiments (LHC) with current experiments (RHIC, Run 2, BaBar)

Data Volume per experiment per year (in units of 10^9 bytes)



Future HEP Experiments

Run 2b at Fermilab

- **Run 2b will start in 2004 and will increase the integrated luminosity to CDF and D0 by a factor of approximately 8 (or more if possible).**
- **It is likely that the computing required will increase by the same factor, in order to pursue the physics topics of interest:**
  - B physics
  - Electroweak
  - Top
  - Higgs
  - Supersymmetry
  - QCD
  - Etc.

Run 2b Computing

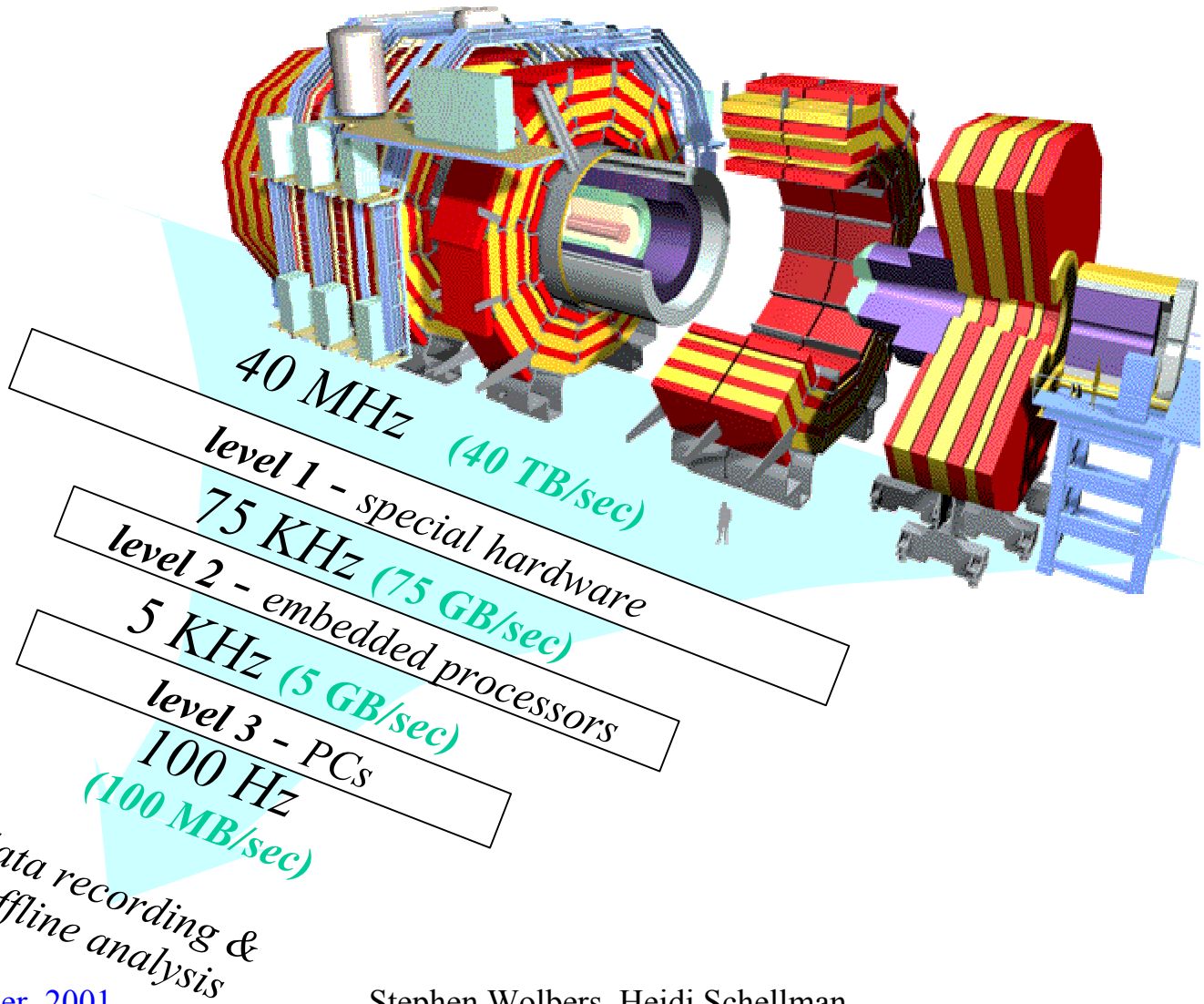
- **Current estimates for Run 2b computing:**
  - 8x CPU, disk, tape storage.
  - Expected cost is same as Run 2a because of increased price/performance of CPU, disk, tape.
  - Plans for R&D testing, upgrades/acquisitions will start next year.
- **Data-taking rate:**
  - May be as large as 100 Mbyte/s (or greater).
  - About 1 Petabyte/year to storage.

Run 2b Computing

- **To satisfy Run 2b Computing Needs:**
  - More CPU (mostly PCs)
  - More Data Storage (higher density tapes)
  - Faster Networks (10 Gbit Ethernet)
  - More Disk
  - More Distributed Computing (GRID)

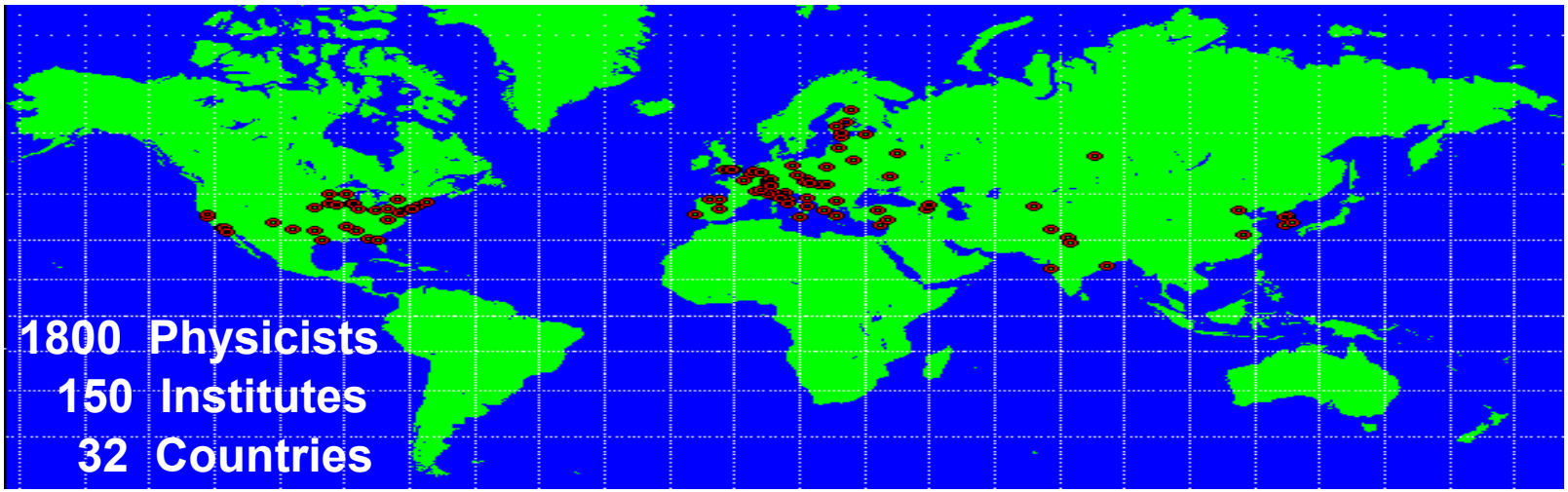
LHC Computing

- LHC (Large Hadron Collider) will begin taking data in 2006-2007 at CERN.
- Data rates per experiment of >100 Mbytes/sec.
- >1 Pbyte/year of storage for raw data per experiment.
- World-wide collaborations and analysis.
  - Desirable to share computing and analysis throughout the world.
  - GRID computing may provide the tools.


40 MHz (40 TB/sec)
level 1 - special hardware
75 KHz (75 GB/sec)
level 2 - embedded processors
5 KHz (5 GB/sec)
level 3 - PCs
100 Hz
(100 MB/sec)
data recording &
offline analysis

CMS Computing Challenges

- Experiment in preparation at CERN/Switzerland
- Strong US participation: ~20%
- Startup: by 2005/2006, will run for 15+ years



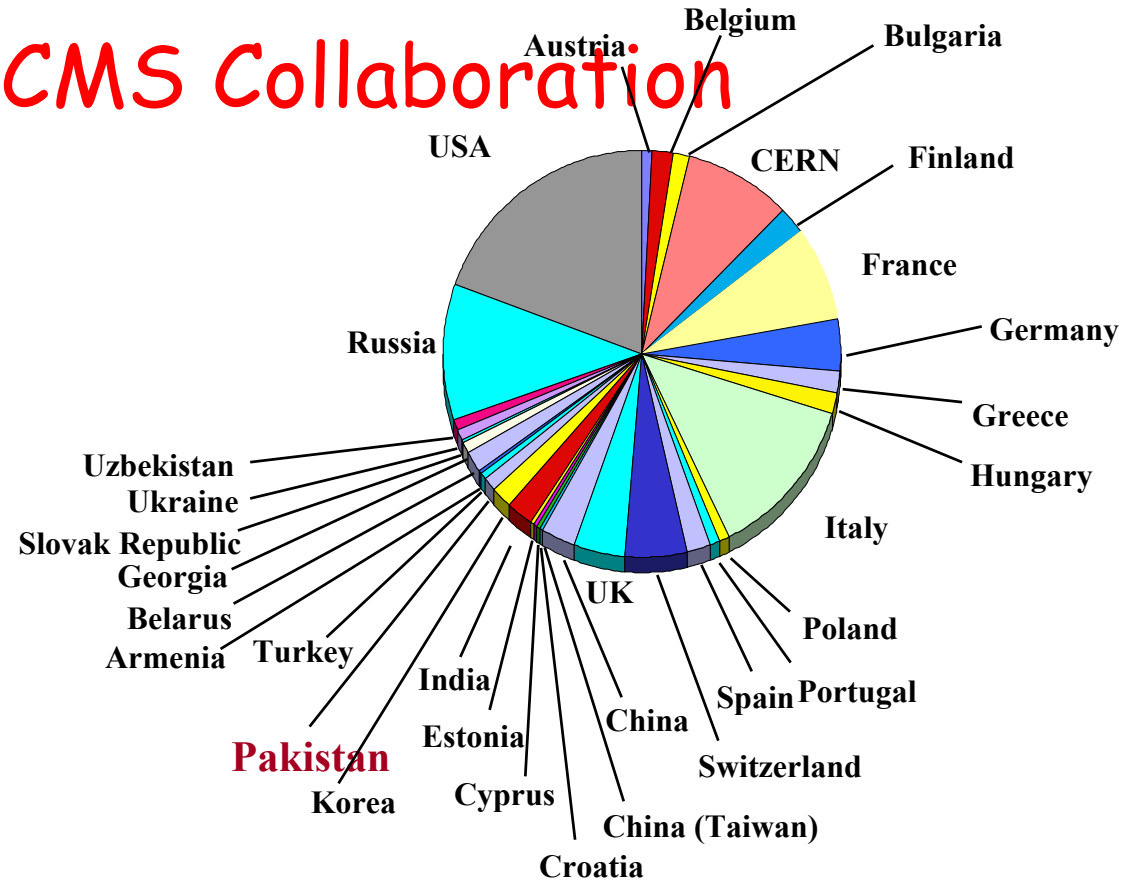
Major challenges associated with:
Communication and collaboration at a distance
Distributed computing resources
Remote software development and physics analysis
R&D: New Forms of Distributed Systems

The CMS Collaboration

	Number of Laboratories
Member States	58
Non-Member States	50
USA	36
Total	144

	Number of Scientists
Member States	1010
Non-Member States	448
USA	351
Total	1809

Associated Institutes	
Number of Scientists	36
Number of Laboratories	5

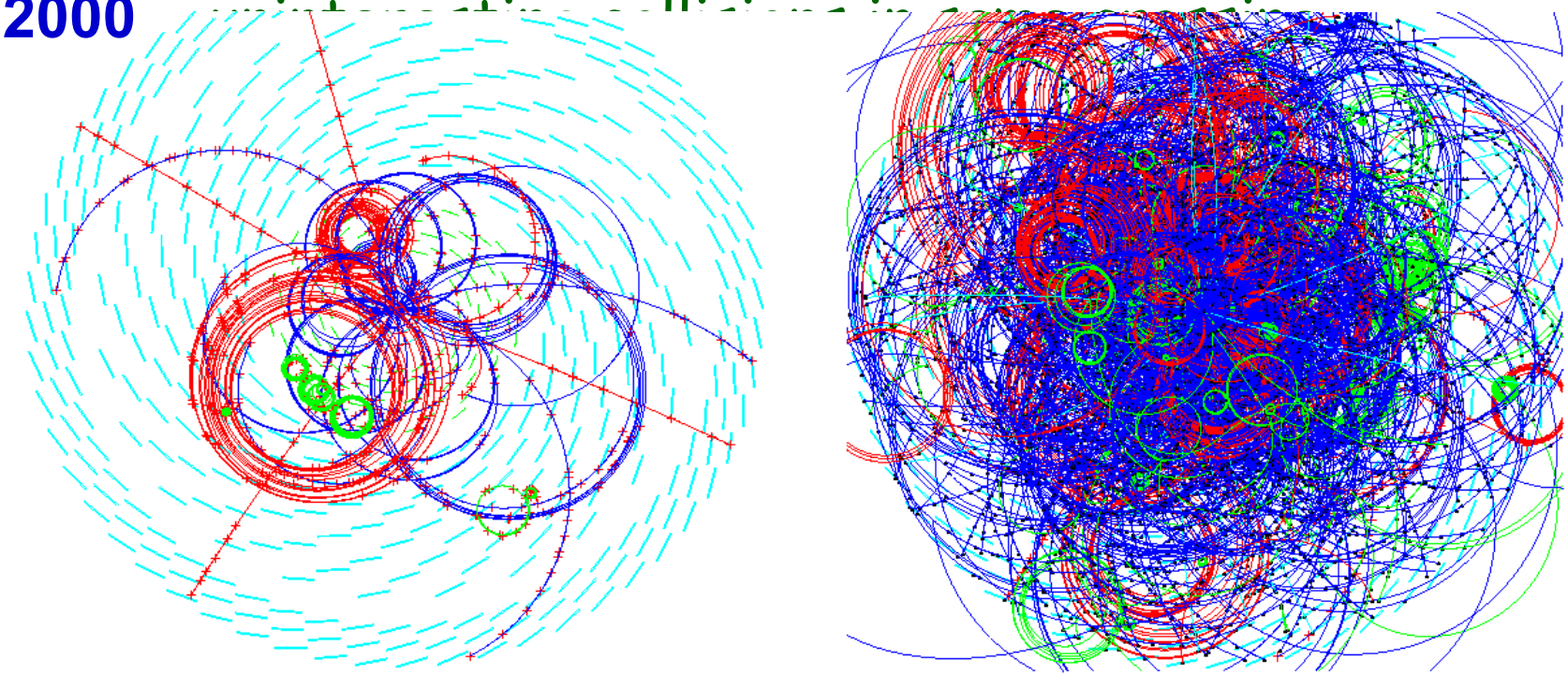


1809 Physicists and Engineers
31 Countries
144 Institutions

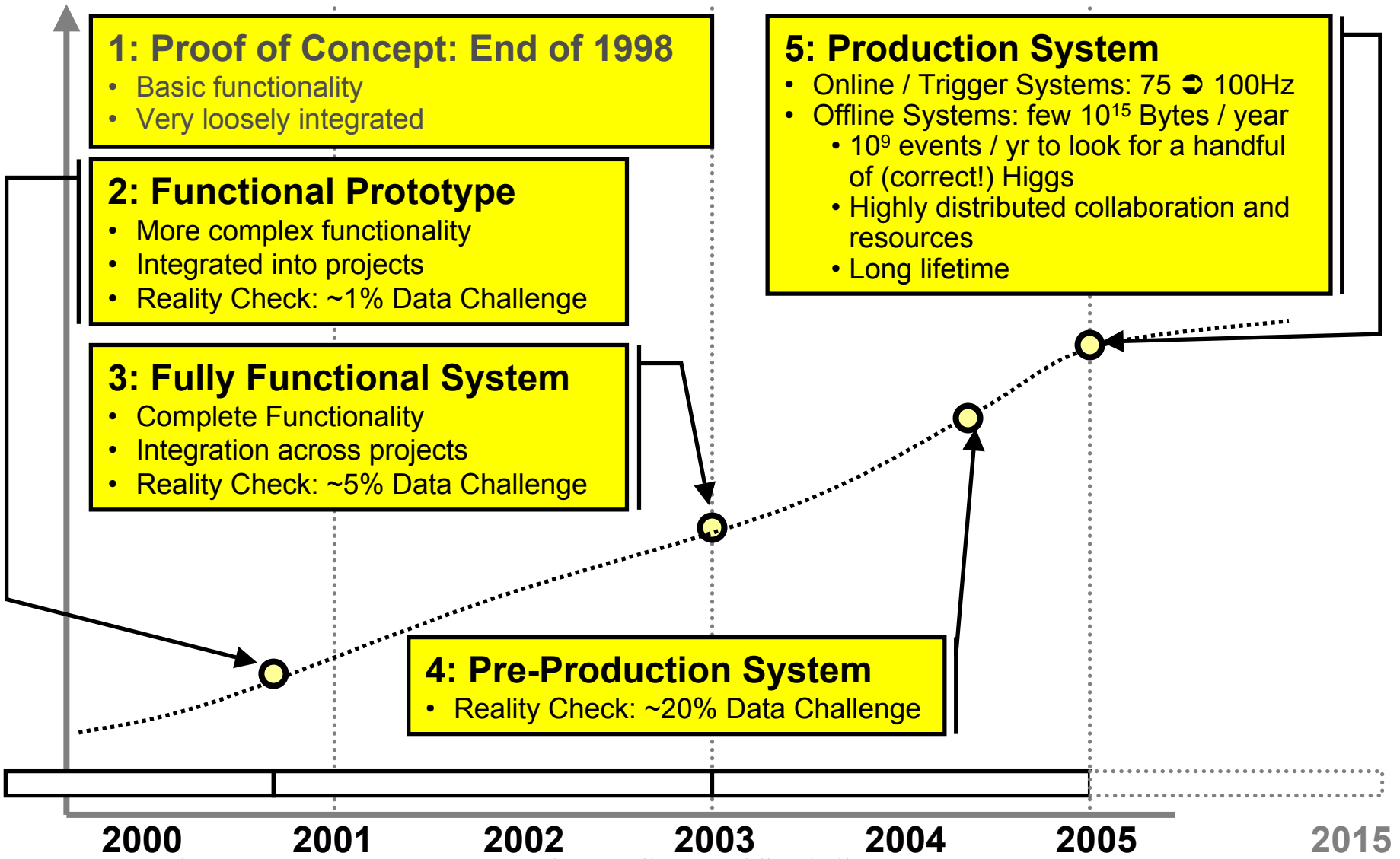
LHC Data Complexity

- **"Events" resulting from beam-beam collisions:**
  - Signal event is obscured by 20 overlapping

2000 **2007**



Software Development Phases



Other Future Experiments

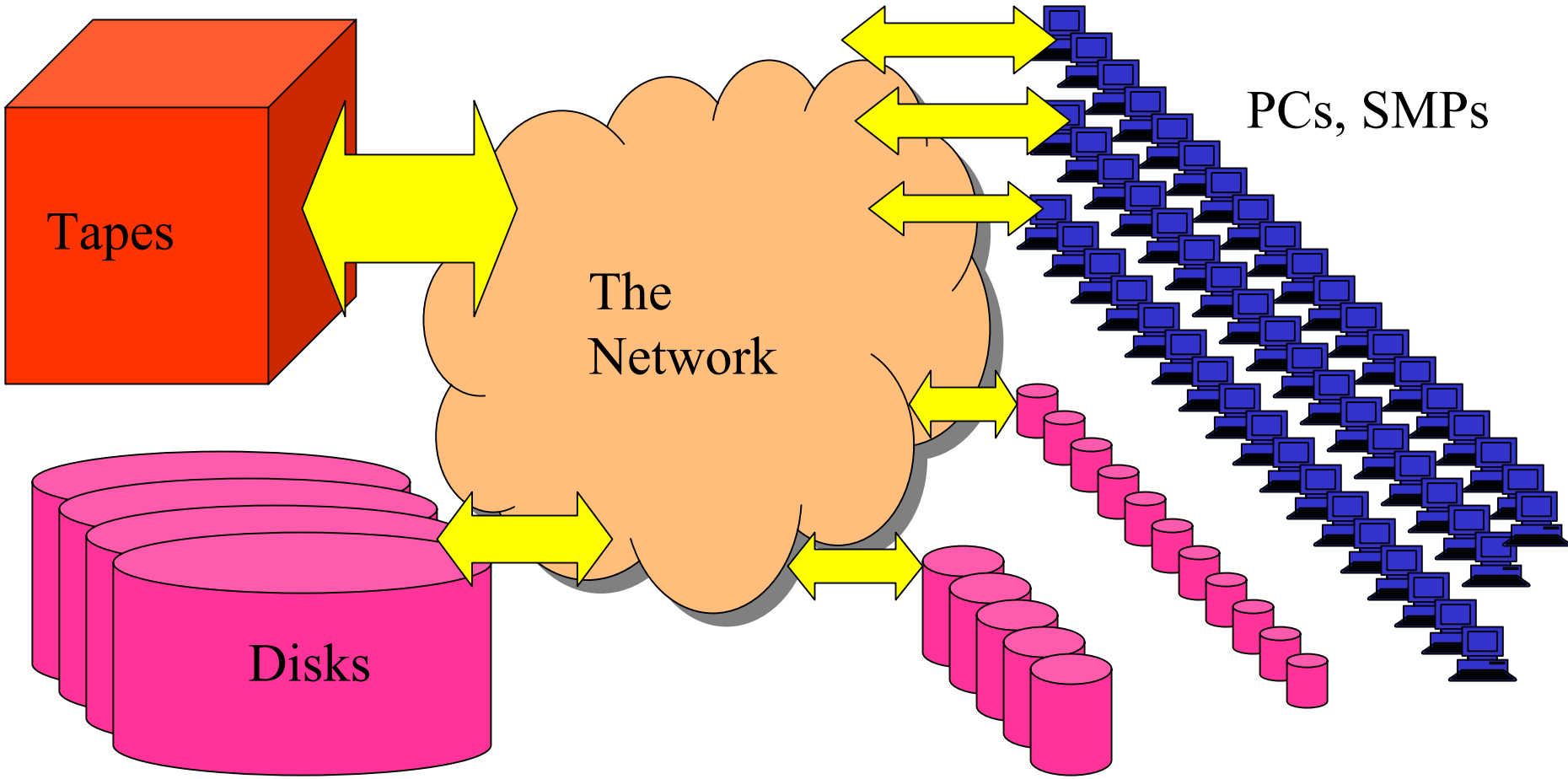
- BaBar, RHIC, JLAB, etc. all have upgrade plans.
- Also new experiments such as BTeV and CKM at Fermilab have large data-taking rates.
- All tend to reach 100 MB/s raw data recording rates during the 2005-2010 timeframe.
- Computing Systems will have to be built to handle the load.

Technology

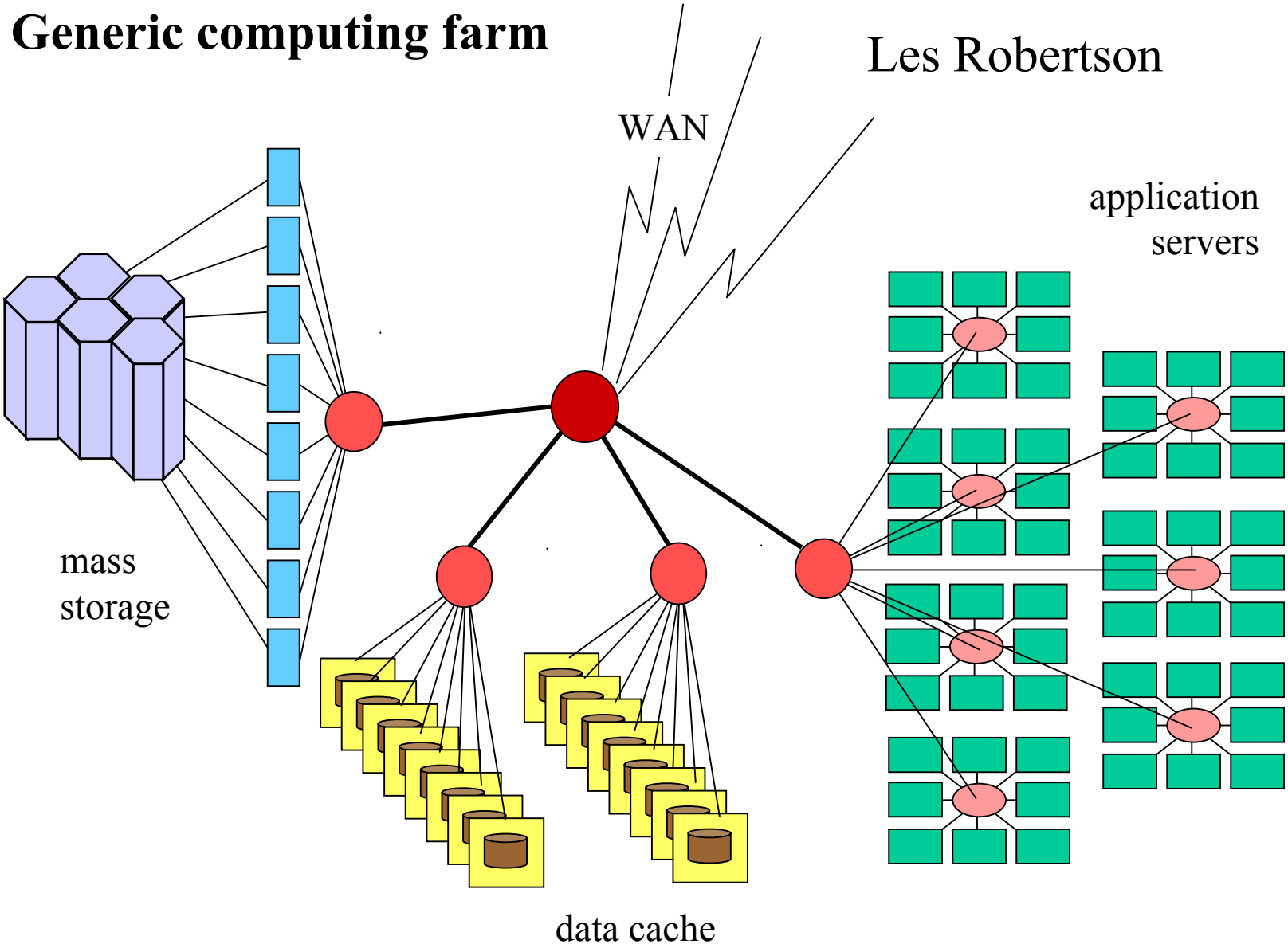
CPU/PCs

- **Commodity Computing has a great deal to offer.**
  - Cheap CPU.
  - Fast network I/O.
  - Fast Disk I/O.
  - Cheap Disk.
- **Can PCs be the basis of essentially all HEP computing in the future?**

Analysis – a very general model



Generic computing farm



Computing Fabric Management

Les Robertson

Key Issues –

- scale
- efficiency & performance
- resilience – fault tolerance
- cost – acquisition, maintenance, operation
- usability
- security

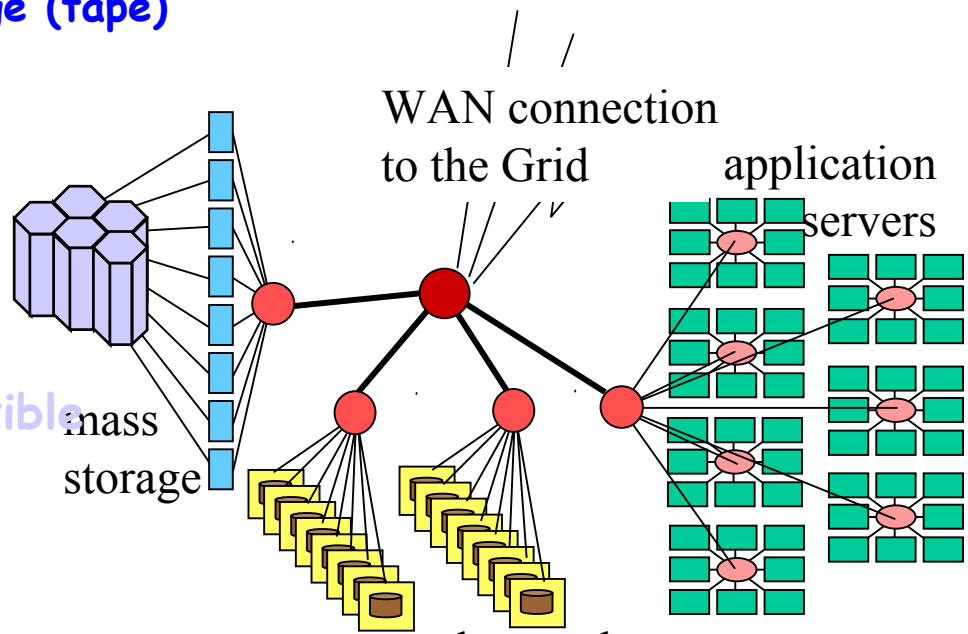
Working assumptions for Computing Fabric at CERN

Les Robertson

- single physical cluster – Tier 0, Tier 1, 4 experiments
  - partitioned by function, (maybe) by user
- an architecture that accommodates mass market components
  and supports cost-effective and seamless capacity evolution
- new level of operational automation
  novel style of fault tolerance – self-healing fabrics
- plan for active mass storage (tape)
  .. but hope to use it only
  as an archive
- one platform –
  Linux, Intel

Where are the industrial products?

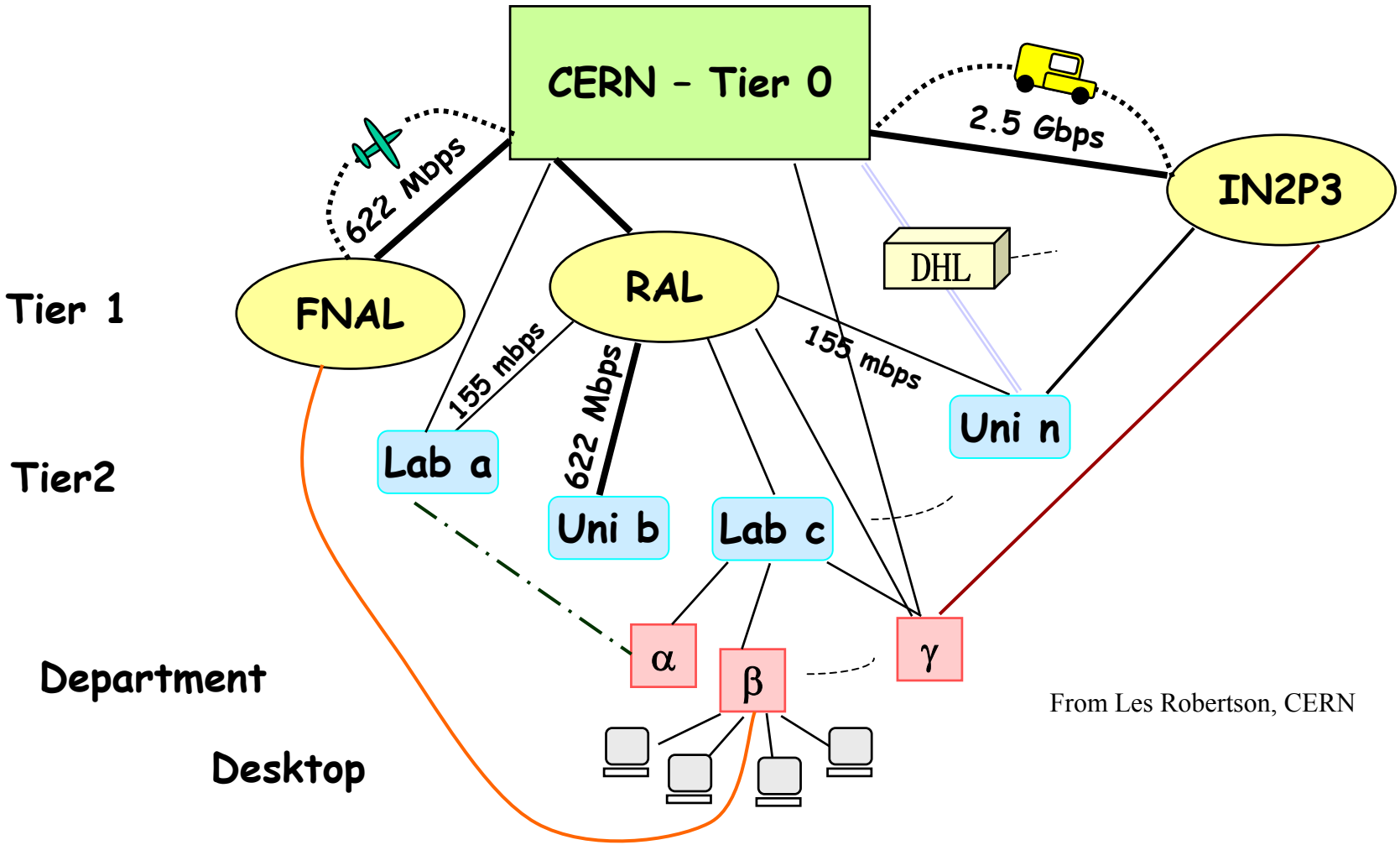
ESSENTIAL to remain flexible on all fronts



GRID Computing

- **GRID Computing has great potential.**
  - Makes use of distributed resources.
  - Allows contributions from many institutions/countries.
  - Provides framework for physics analysis for the future.

CMS/ATLAS and GRID Computing



From Les Robertson, CERN

CERN/Outside Resource Ratio ~1:2
Tier0/(Σ Tier1)/(Σ Tier2) ~1:1:1

Example: CMS Data Grid

Experiment

~PBytes/sec

Online System

~100 MBytes/sec

Bunch crossing per 25 nsecs.
100 triggers per second
Event is ~1 MByte in size

Tier 0 +1

CERN Computer Center > 20 TIPS

2.5 Gbits/sec

France Center

UK Center

Italy Center

USA Center

Tier 1

2.5 Gbits/sec

Tier 2

Tier2 Center

~622 Mbits/sec

Tier 3

Institute ~0.25TIPS

Physics data cache

100 - 1000 Mbits/sec

Tier 4

Workstations, other portals

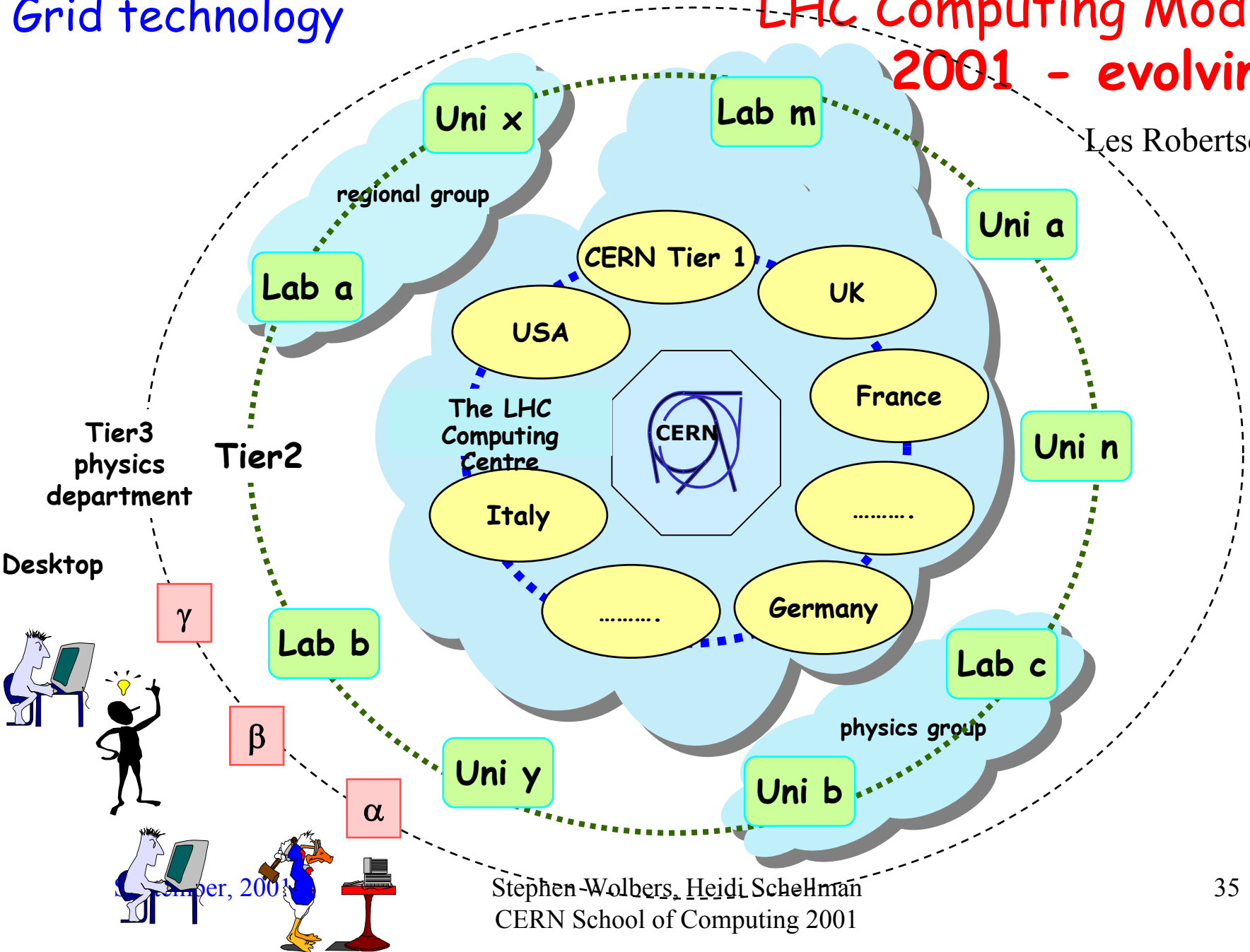
Physicists work on analysis "channels".

Each institute has ~10 physicists working on one or more channels

The opportunity of Grid technology

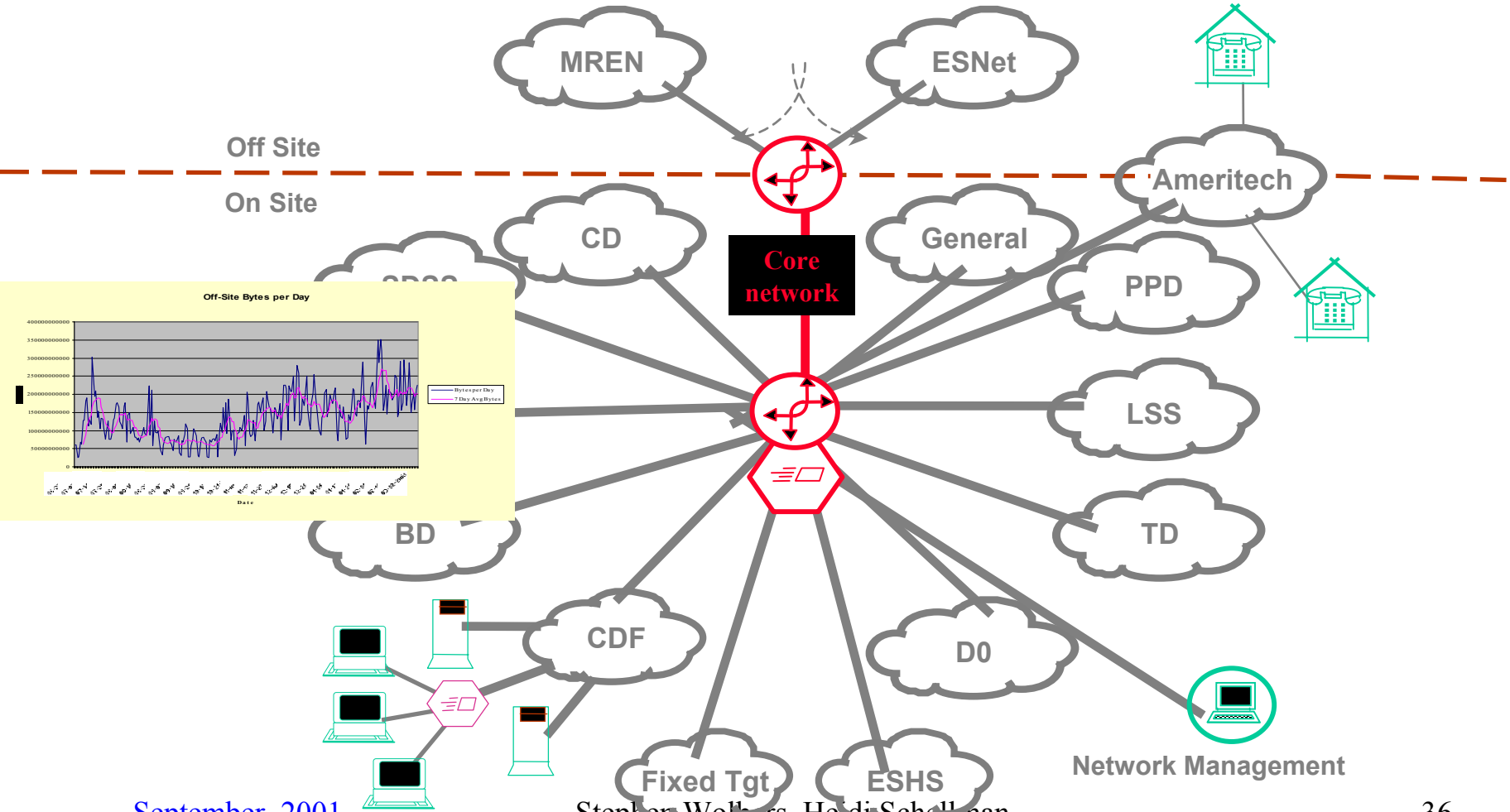
LHC Computing Model 2001 - evolving

Les Robertson



les.robertson@cern.ch

Fermilab Networking and connection to Internet



Are Grids a solution?

Computational Grids Les Robertson, CERN

- **Change of orientation of Meta-computing activity**
  - From inter-connected super-computers … .. towards a more general concept of a computational power Grid (The Grid – Ian Foster, Carl Kesselman**)
- **Has found resonance with the press, funding agencies**

But what is a Grid?

*"Dependable, consistent, pervasive access to resources**"*

So, in some way Grid technology makes it easy to use diverse, geographically distributed, locally managed and controlled computing facilities – as if they formed a coherent local cluster

** Ian Foster and Carl Kesselman, editors, "The Grid: Blueprint for a New Computing Infrastructure," Morgan Kaufmann, 1999

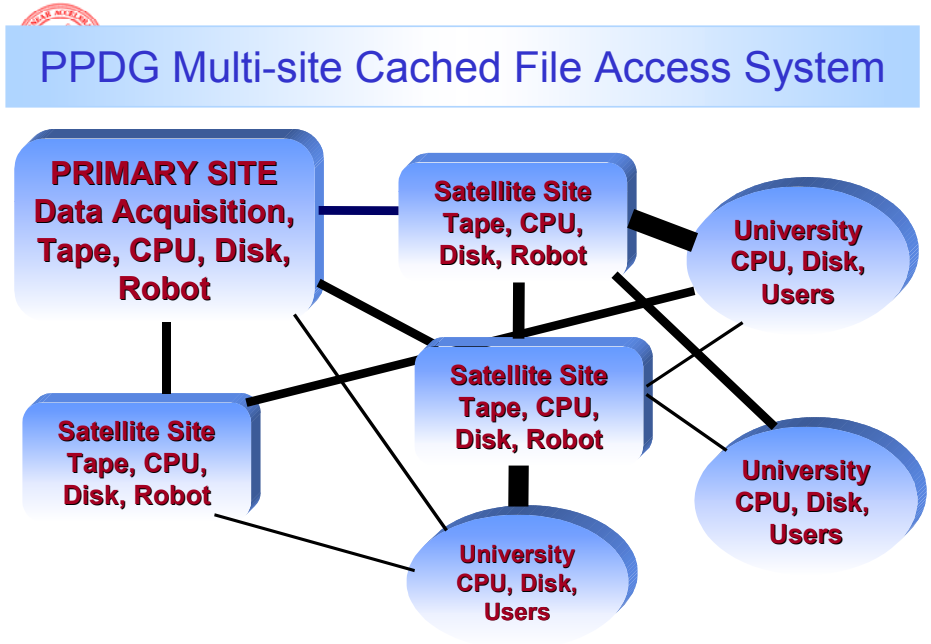
What does the Grid do for you?

Les Robertson

- **You submit your work**
- **And the Grid**
  - Finds convenient places for it to be run
  - Organises efficient access to your data
    - Caching, migration, replication
  - Deals with authentication to the different sites that you will be using
  - Interfaces to local site resource allocation mechanisms, policies
  - Runs your jobs
  - Monitors progress
  - Recovers from problems
  - Tells you when your work is complete
- **If there is scope for parallelism, it can also decompose your work into convenient execution units based on the available resources, data distribution**

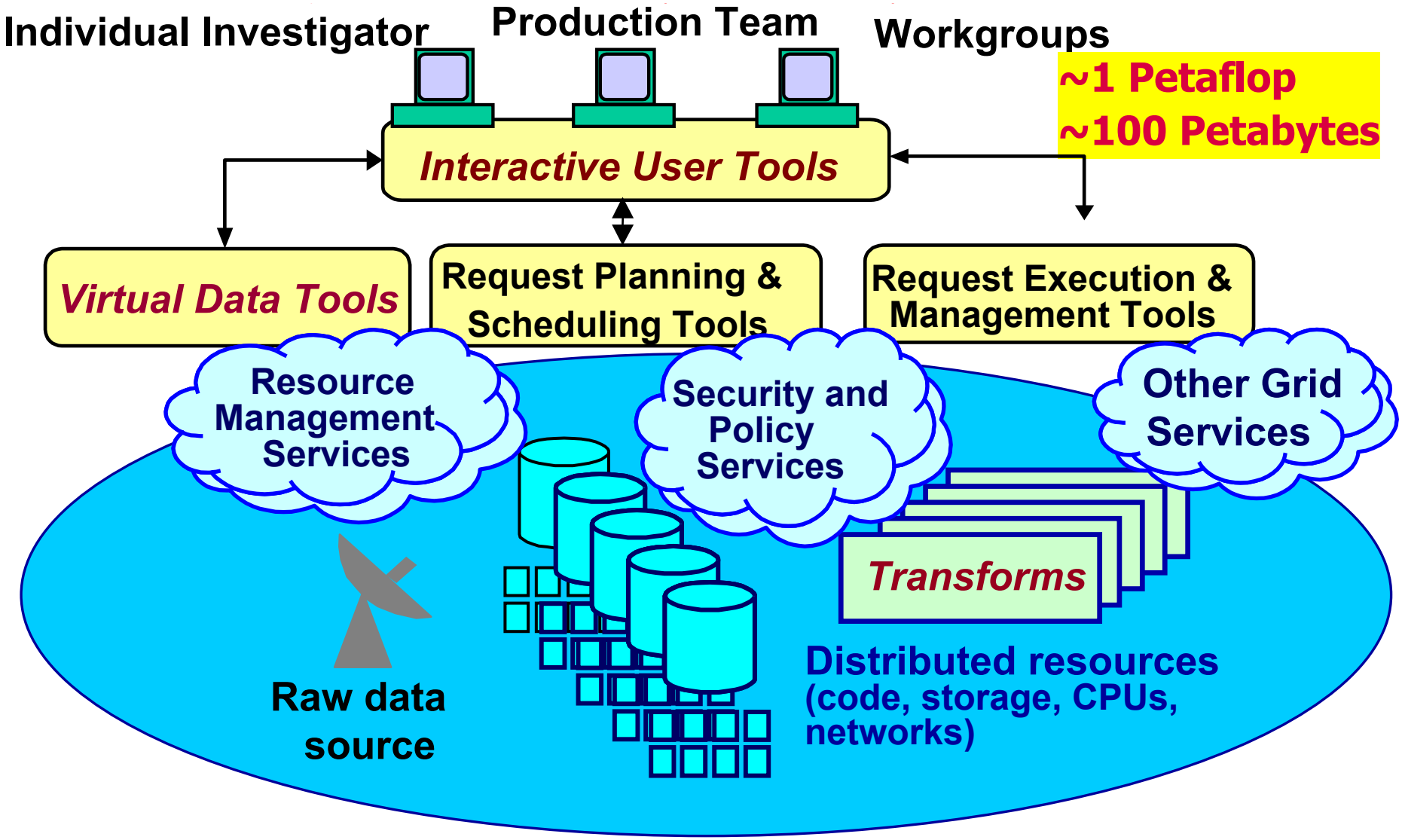
PPDG GRID R&D

Richard Mount, SLAC



GriPhyN Overview (www.griphyn.org)

- **5-year, $12M NSF ITR proposal to realize the concept of virtual data, via:**
  1) CS research on
     - Virtual data technologies (info models, management of virtual data software, etc.)
     - Request planning and scheduling (including policy representation and enforcement)
     - Task execution (including agent computing, fault management, etc.)
  2) Development of Virtual Data Toolkit (VDT)
  3) Applications: ATLAS, CMS, LIGO, SDSS
- **PIs=Avery (Florida), Foster (Chicago)**


Individual Investigator
Production Team
Workgroups
~1 Petaflop
~100 Petabytes
Interactive User Tools
Virtual Data Tools
Request Planning & Scheduling Tools
Request Execution & Management Tools
Resource Management Services
Security and Policy Services
Other Grid Services
Transforms
Raw data source
Distributed resources (code, storage, CPUs, networks)

Carl Kesselman
Center for Grid Technologies
USC/Information Sciences Institute

Globus Applications and Deployments

- **Application projects include**
  - GriPhyN, PPDG, NEES, EU DataGrid, ESG, Fusion Collaboratory, etc., etc.
- **Infrastructure deployments include**
  - DISCOM, NASA IPG, NSF TeraGrid, DOE Science Grid, EU DataGrid, etc., etc.
  - UK Grid Center, U.S. GRIDS Center
- **Technology projects include**
  - Data Grids, Access Grid, Portals, CORBA, MPICH-G2, Condor-G, GrADS, etc., etc.

Carl Kesselman
Center for Grid Technologies
USC/Information Sciences Institute

Example Application Projects

- AstroGrid: astronomy, etc. (UK)
- Earth Systems Grid: environment (US DOE)
- EU DataGrid: physics, environment, etc. (EU)
- EuroGrid: various (EU)
- Fusion Collaboratory (US DOE)
- GridLab: astrophysics, etc. (EU)
- Grid Physics Network (US NSF)
- MetaNEOS: numerical optimization (US NSF)
- NEESgrid: civil engineering (US NSF)
- Particle Physics Data Grid (US DOE)

HEP Related Data Grid Projects

Paul Avery

- **Funded projects**
  - GriPhyN — USA — NSF, $11.9M + $1.6M
  - PPDG I — USA — DOE, $2M
  - PPDG II — USA — DOE, $9.5M
  - EU DataGrid — EU — $9.3M
- **Proposed projects**
  - iVDGL — USA — NSF, $15M + $1.8M + UK
  - DTF — USA — NSF, $45M + $4M/yr
  - DataTag — EU — EC, $2M?
  - GridPP — UK — PPARC, > $15M
- **Other national projects**
  - UK e-Science (> $100M for 2001-2004)
  - Italy, France, (Japan?)

GRID Computing

- GRID computing is a very hot topic at the moment.
- HENP is involved in many GRID R&D projects, with the next steps aimed at providing real tools and software to experiments.
- The problem is a large one and it is not yet clear that the concepts will turned into effective computing.
  - CMS@HOME?

The full costs?

Matthias Kasemann

- Space
- Power, cooling
- Software

- LAN
- Replacement/Expansion 30% per year

- Mass storage

- People

Storing Petabytes of Data in mass storage

- **Storing (safely) petabytes of data is not easy or cheap.**
  - Need large robots (for storage and tape mounting).
  - Need many tapedrives to get the necessary I/O rates.
    - Tapedrives and tapes are an important part of the solution, and has caused some difficulty for Run 2.
  - Need bandwidth to the final application (network or SCSI).
  - Need system to keep track of what is going on and schedule and prioritize requests.

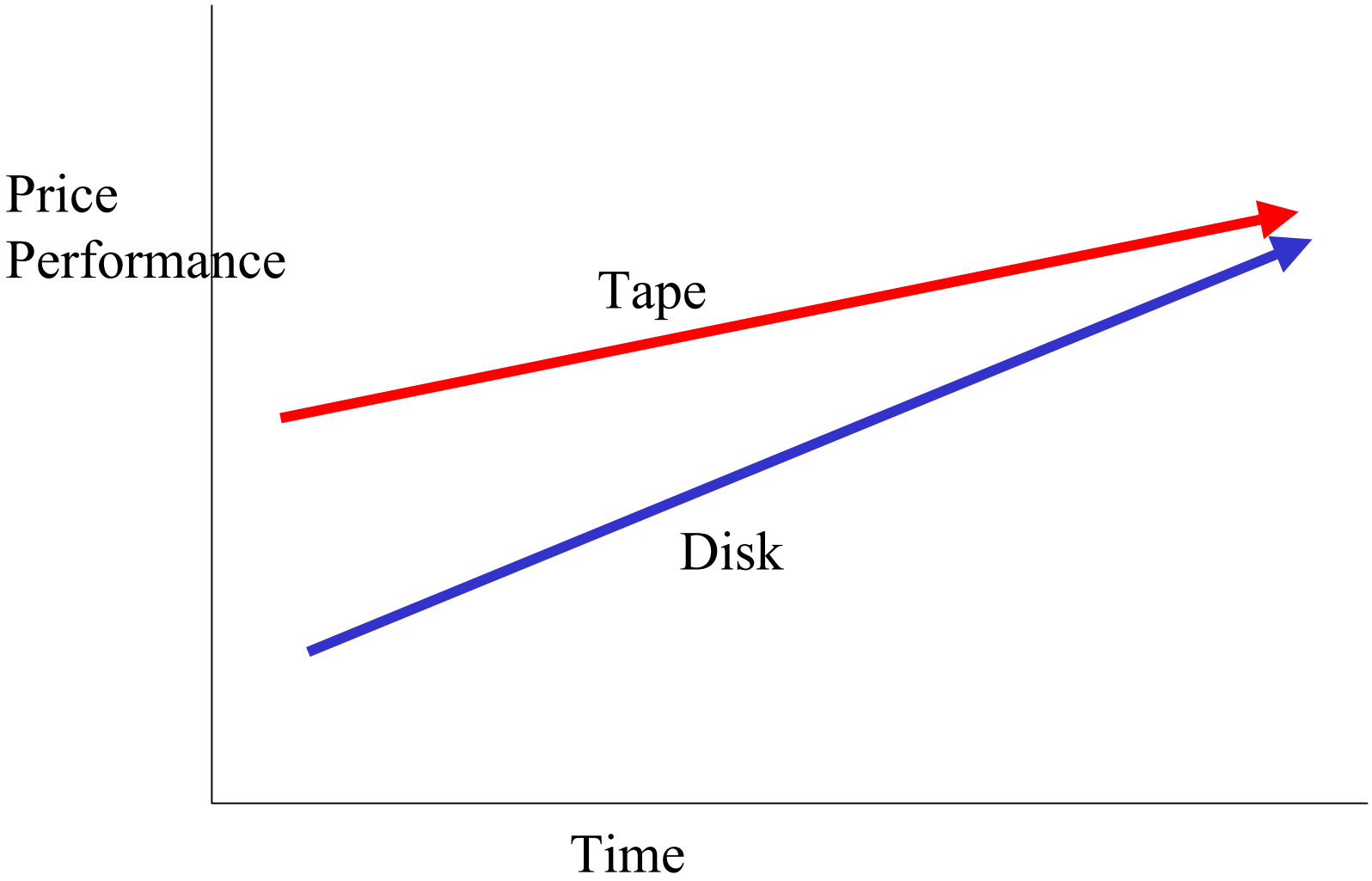
Tapedrives and tapes

- Tapedrives are not always reliable, especially when one is pushing for higher performance at lower cost.
- Run 2 choice was Exabyte Mammoth 2.
  - 60 Gbytes/tape.
  - 12 Mbyte/sec read/write speed.
  - About $1 per Gbyte for tape. (A lot of money.)
  - $5000 per tapedrive.
- Mammoth 2 was not capable (various problems).
- AIT2 from SONY is the backup solution and is being used by CDF.
- STK 9940 was chosen by D0 for data, LTO for Monte Carlo.
- Given the Run 2 timescale, upgrades to newer technology will occur.
- Finally, Fermilab is starting to look at PC diskfarms to replace tape completely.

Robots and tapes

Disk Farms (Tape Killer)

- **Tapes are a pain:**
  - They are slow
  - They wear out and break
  - They improve ever so slowly
- **But they have advantages:**
  - Large volume of data
  - Low price
  - Archival medium


Price
Performance
Tape
Disk
Time

An Idea: Disk Farms

- **Can we eliminate tape completely for data storage?**
- **What makes this possible?**
  - Disk drives are fast, cheap, and large.
  - Disk drives are getting faster, cheaper and larger.
  - Access to the data can be made via the standard network-based techniques
    - NFS,AFS,tcp/ip,fibrechannel
  - Cataloging of the data can be similar to tape cataloging

Disk Farms

- **Two Ideas:**
  - Utilize disk storage on cheap PCs
  - Build storage devices to replace tape storage
- **Why Bother?**
  - The price performance of disk is increasing very rapidly.
  - Tape performance is not improving as quickly.

I.-Utilize cheap disks on PCs

- **All PCs come with substantial EIDE disk storage**
  - Cheap
  - Fast
  - On CPU farms it is mostly unused
- **Given the speed of modern ethernet switches, this disk storage can be quite useful**
  - Good place to store intermediate results
  - Could be used to build a reasonable performance SAN

II.-Build a true disk-based mass storage system

- **Components of all-disk mass storage:**
  - Large number of disks.
  - Connected to many PCs.
  - Software catalog to keep track of files.
- **Issues**
  - Power, cooling.
  - Spin-down disks when not used?
  - Catalog and access

Summary of Lecture 3

- Future HEP experiments require massive amounts of computing, including data collection and storage, data access, database access, computing cycles, etc.
- Tools for providing those cycles exist, and an architecture for each experiment needs to be invented.
- The GRID will be a part of this architecture and is an exciting prospect to help HEP.